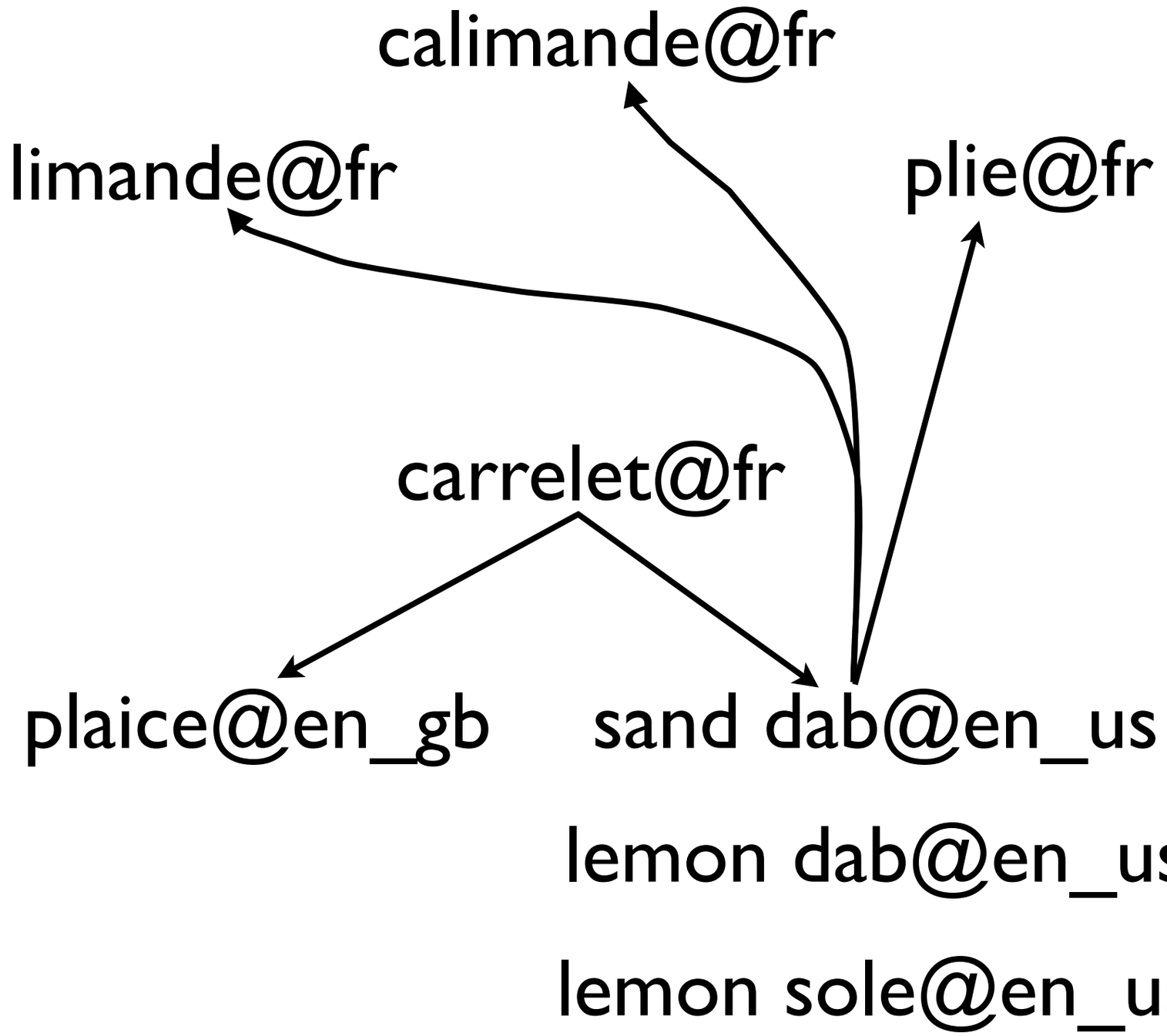
Linking Open Government Data

David Wood

14 March 2012

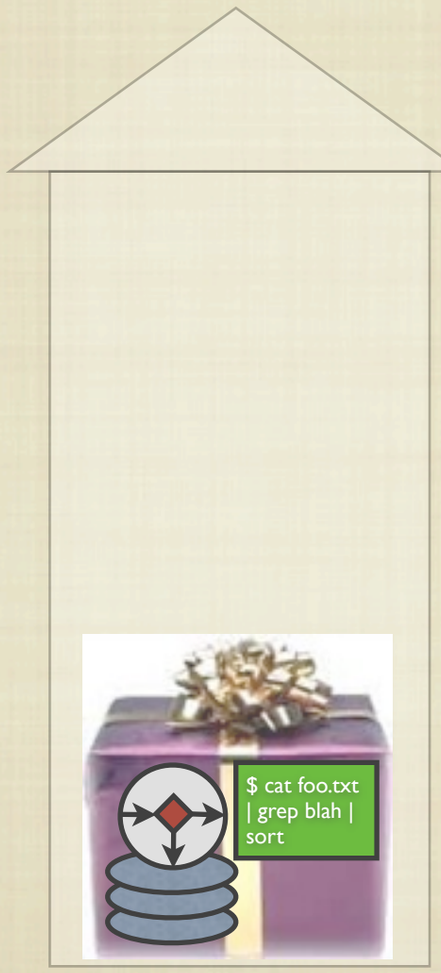
NASA Goddard

http://purl.org/net/prototypo/logd



Julia Child had a problem. Lack of governmental data standards hampered her work on her 1961 bestseller Mastering the Art of French Cooking. Names for fish varied between British & American English and French. Even Latin names differed. French and US governments were working on these problems themselves. The data was in **silos.**

Silos

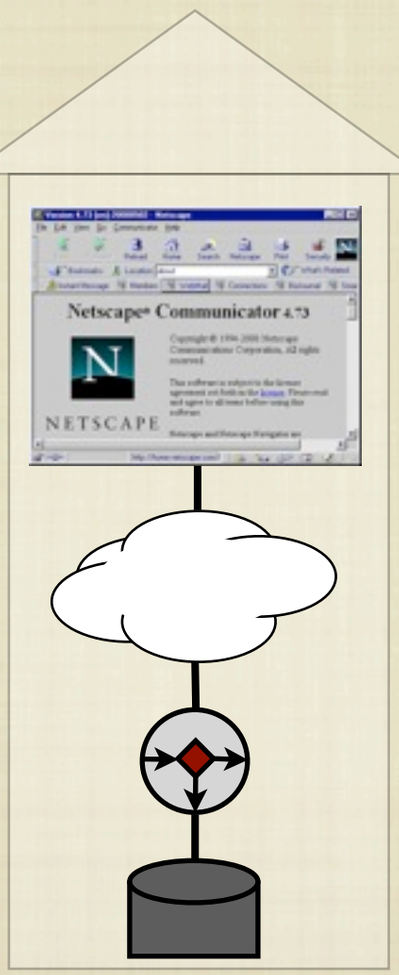


1970s

A neat little package

1980s

Client-Server



1990s

The Early Web

Monday, March 12, 12

Julia had her problem because her data was separated into silos. The history of computing is littered with silos.

Silos

Email

Databases

Proprietary containers

Monday, March 12, 12

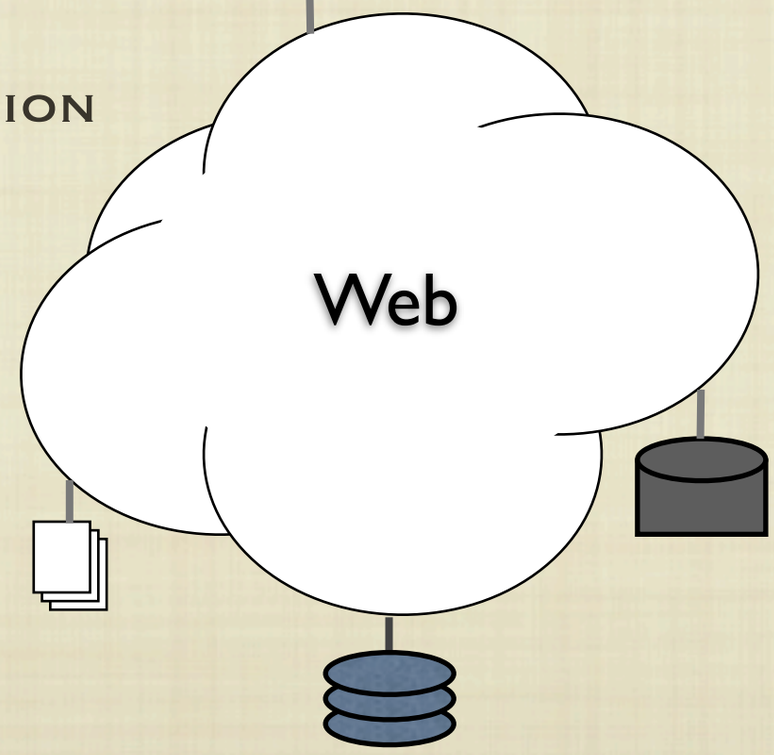
We make new silos every day. There are lots of ways to make silos.

Universal Client

Ubiquitous,
reusable applications

URL
Curation/
Proxy/Cache

Universal Connection



Universal File
System

The Web is very different. Documents and processes can link to each other: Facebook, twitter, Wordpress, Drupal, even Salesforce and Google Docs can break out of their silos via hyperlinks.

The Web of Documents

- A global network of linked documents
- A place where anyone can say anything about anything
- A vast collection of human-readable knowledge (and opinion)
- Documents are linked, but links are not qualified

Data Inflation

Megabyte (MB) = 2^{20}

Gigabyte (GB) = 2^{30}

Terabyte (TB) = 2^{40} 1,000GB

Petabyte (PB) = 2^{50} 1,000TB

Exabyte (EB) = 2^{60} or 1,000PB

Zettabyte (ZB) = 2^{70} or 1,000EB

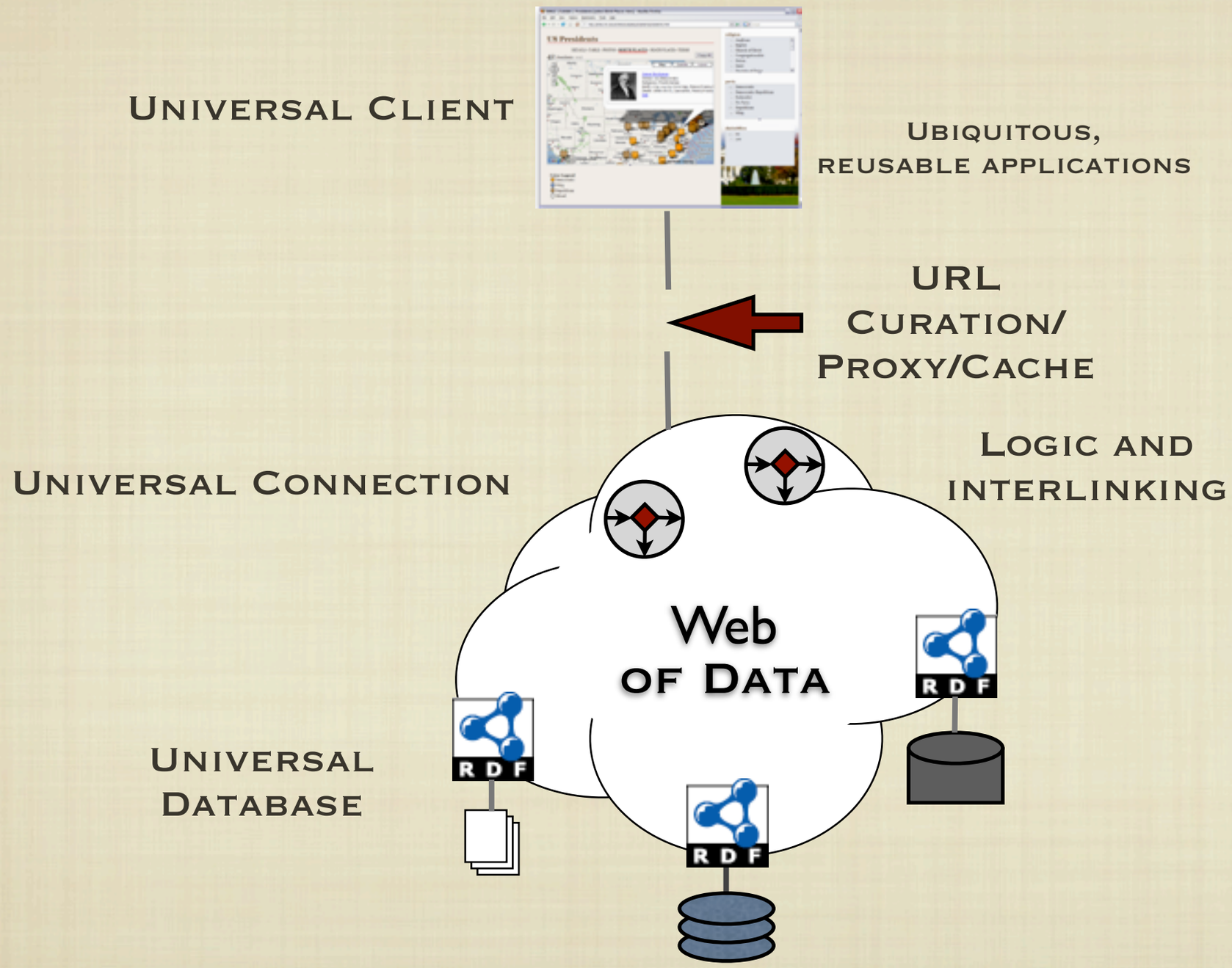
Yottabyte (YB) = 2^{80} or1,000ZB

"We are beginning to routinely deal with vast quantities of data and only through a metadata management strategy can we tackle the quantity of data to look for the nuggets for gold."

Credit: The Economist, Monstrous Amounts of Data, Feb 27th, 2010 Special Report on Managing Information.

Our data is still in the dinosaur age. Traditional data is hierarchical, tabular with external schemas, and so are the systems that support them.

Photo credit: David Wood, 2009



Monday, March 12, 12

By using the Web to interlink structured data, we can remove data silos like we removed file and process silos. We can make the Web into a database - a very large, distributed database - which will be as different from the RDBMS as early software is from the Web.

The Web of Data

- A global network of linked **statements**
- A place where anyone can say anything about anything
- A vast collection of **machine**-readable knowledge (and opinion)
- **Statements** are linked, and links **are qualified**

National Aeronautics and Space Administration

Thursday, 22 April 2010

NSSDC Master Catalog Search

- + Spacecraft
- + Experiments
- + Data Collections
- + Personnel
- + Publications
- + Maps
- + New/Updated Data
- + Lunar/Planetary Events

Apollo 11 Command and Service Module (CSM)

National Space Science Data Center

Apollo 11 Command and Service Module (CSM)

NSSDC ID: 1969-059A

Description

Apollo 11 was the first mission in which humans walked on the lunar surface and returned to Earth. On 20 July 1969 two astronauts (Apollo 11 Commander Neil A. Armstrong and LM pilot Edwin E. "Buzz" Aldrin Jr.) landed in Mare Tranquilitatis (the Sea of Tranquility) on the Moon in the Lunar Module (LM) while the Command and Service Module (CSM) (with CM pilot Michael Collins) continued in lunar orbit. During their stay on the Moon, the astronauts set up scientific experiments, took photographs, and collected lunar samples. The LM took off from the Moon on 21 July and the astronauts returned to Earth on 24 July.

Mission Profile
After launch on Saturn V SA-504 on 16 July 1969 at 13:32 UT (9:32 a.m. EDT) from pad 39A of Kennedy Space Center, Apollo 11 entered Earth orbit. After 1 1/2 Earth orbits, the S-IVB stage was re-ignited at 16:16:16 UT for a translunar injection burn of 5 minutes, 48 seconds putting the spacecraft on course for the Moon. The CSM separated from the S-IVB stage containing the LM 33 minutes later, turned around and docked with the LM at 16:56:03 UT. About an hour and 15 minutes later the S-IVB stage was injected into heliocentric orbit. During translunar coast a color TV transmission was made from Apollo 11 and on 17 July a 3-second mid-course correction burn of the main engine was performed. Lunar orbit insertion was achieved on 19 July at 17:21:50 UT by a retrograde firing of the main engine for 357.5 seconds while the spacecraft was behind the Moon and out of contact with Earth. A later 17 second burn circularized the orbit. On 20 July

Alternate Names

- Apollo 11 CSM
- Columbia
- CSM-107
- 04039

Facts in Brief

Launch Date: 1969-07-16
Launch Vehicle: Saturn 5
Launch Site: Cape Canaveral, United States
Mass: 28801.0 kg

Funding Agency

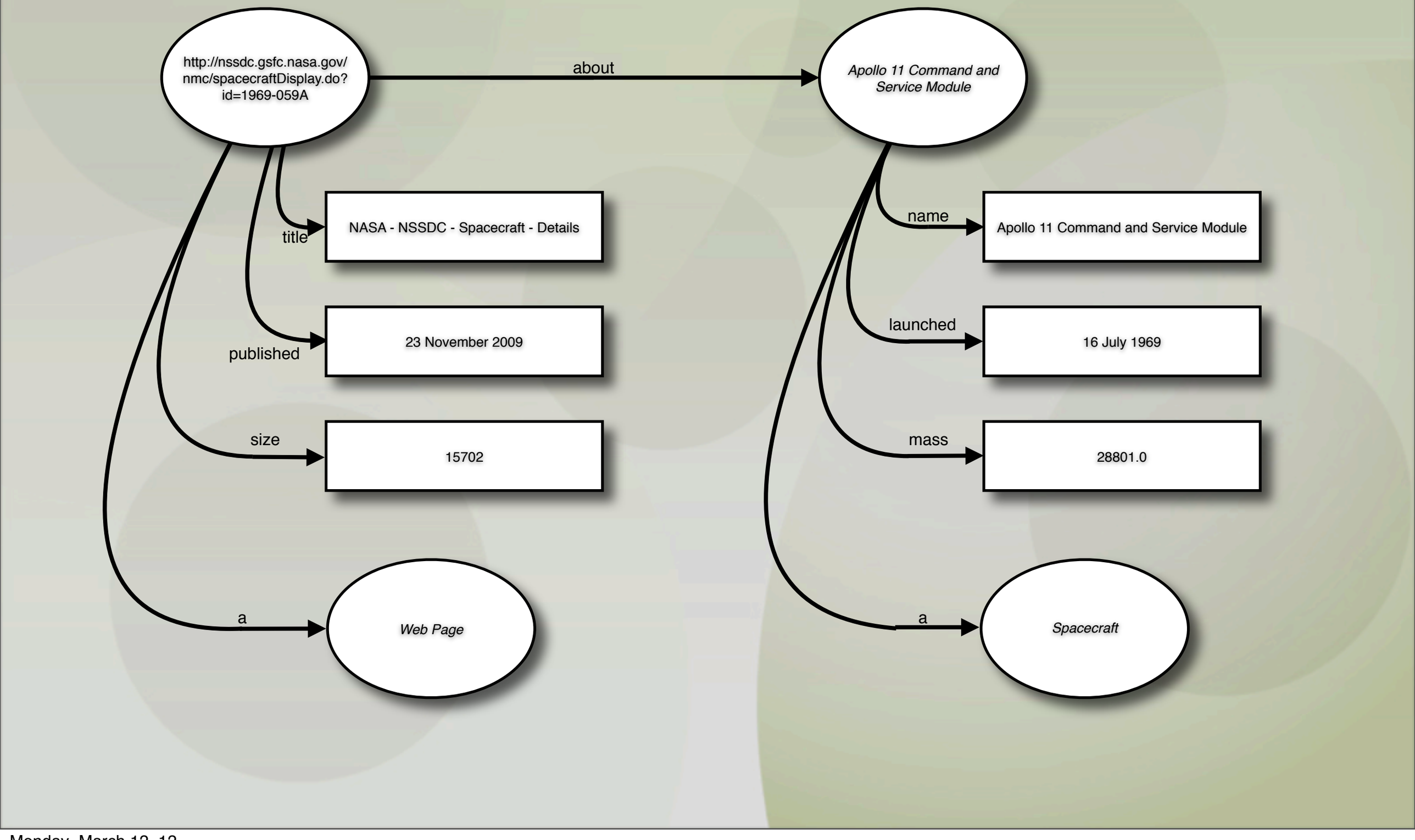
- NASA-Office of Manned Space Flight (United States)

Disciplines

- Human Crew
- Planetary Science

Monday, March 12, 12

What does Linked Data look like? This Web page describes the Apollo 11 CSM.



A Linked Data description simply lists all the resources and connects them with arrows representing properties (or predicates). URIs are assigned to anything that is not a literal (content shown in ovals, and the predicates).

Spacecraft	Name	Mass	Launched	Launch Site
1969-018A	Apollo 9	26801.0	1969-03-03	1
1969-030A	Mariner 7	411.8	1969-03-27	1
1969-043A	Apollo 10	28834.0	1969-05-18	1
1969-059A	Apollo 11 Command and Service Module (CSM)	28801.0	1969-07-16	1
1969-059C	Apollo 11 Lunar Module / EASEP	15065.0	1969-07-16	1
1969-099A	Apollo 12 Command and Service Module (CSM)	28790.0	1969-11-14	1
1970-029C	Apollo 13 Lunar Module/ALSEP	15196.0	1970-04-11	1
1970-029A	Apollo 13 Command and Service Module (CSM)	28945.0	1970-04-11	1

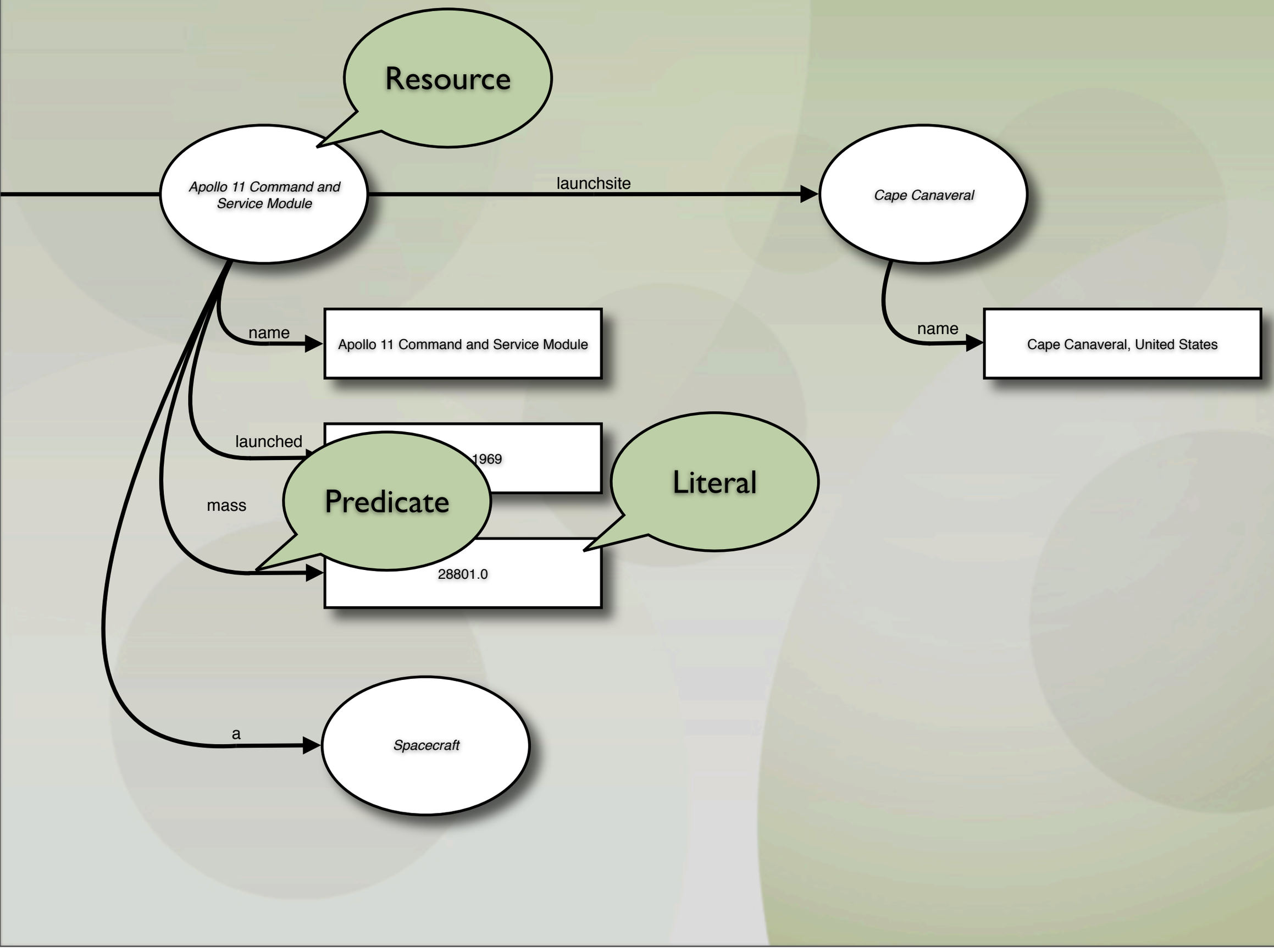
This could be represented this way in an RDBMS.

Spacecraft	Name	Mass	Launched	Launch Site
1969-018A	Apollo 9	26801.0	1969-03-03	1
1969-030A	Mariner 7	411.8	1969-03-27	1
1969-043A	Apollo 10			
1969-059A	Apollo 11 Command and Service M			
1969-059C	Apollo 11 Lunar Module / EA			
1969-099A	Apollo 12 Command and Service M			
1970-029C	Apollo 13 Lunar Module/ALSEP	15196.0	1970-04-11	1
1970-029A	Apollo 13 Command and Service Module (CSM)	28945.0	1970-04-11	1

ID	Name
1	Cape Canaveral, United States
2	Tyuratam (Baikonur Cosmodrome), U.S.S.R
3	Xichang Space Launch Center

Monday, March 12, 12

... but you would need another table for the launch sites.



With Linked Data, we just add another relationship.

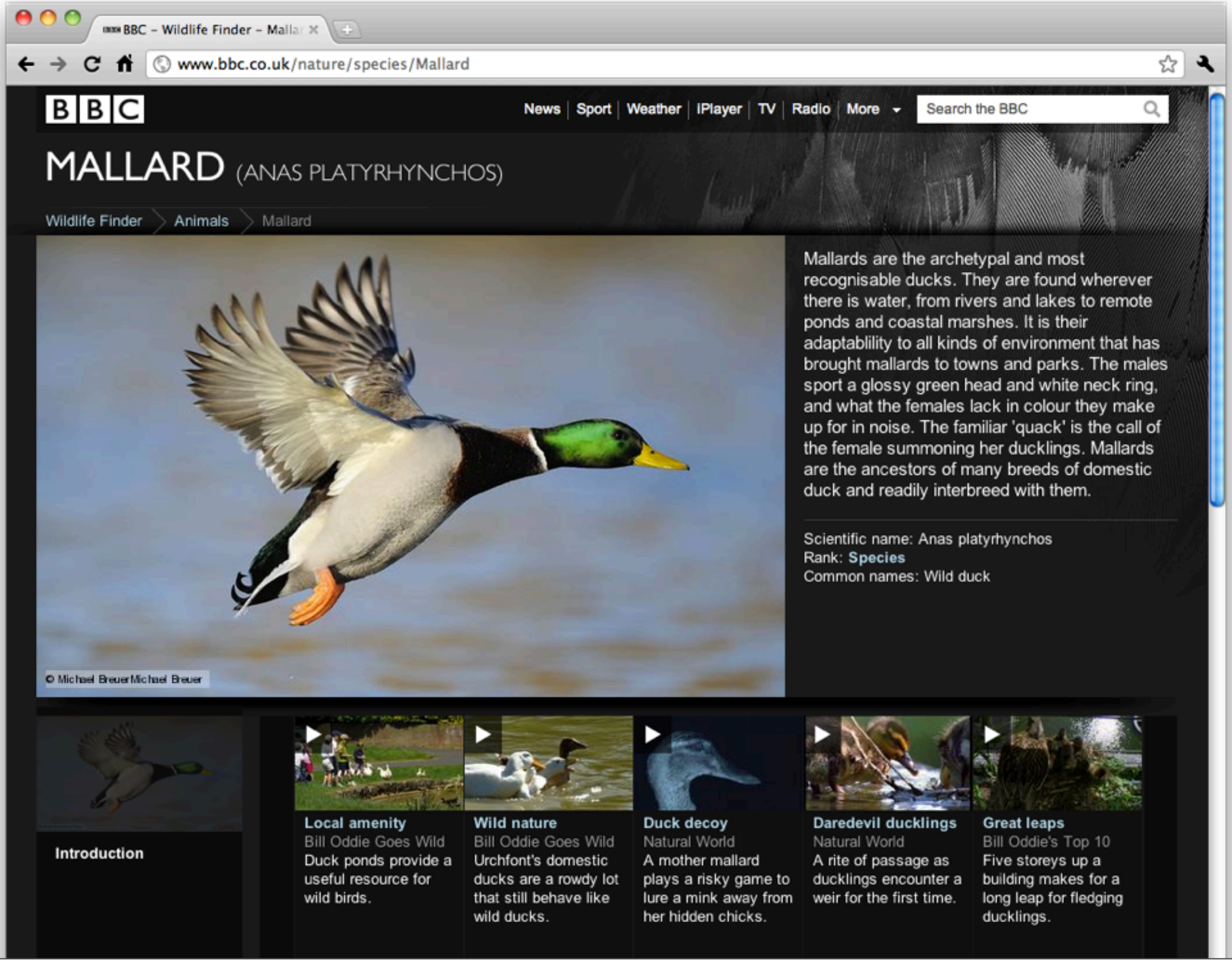
Statement or Triple

Subject Predicate Object

Subject Predicate Object

Statement or Triple

In this way, we create a (potentially large and maybe distributed!) graph out of “triples”, or RDF statements.



Monday, March 12, 12

Mallard Web page, BBC

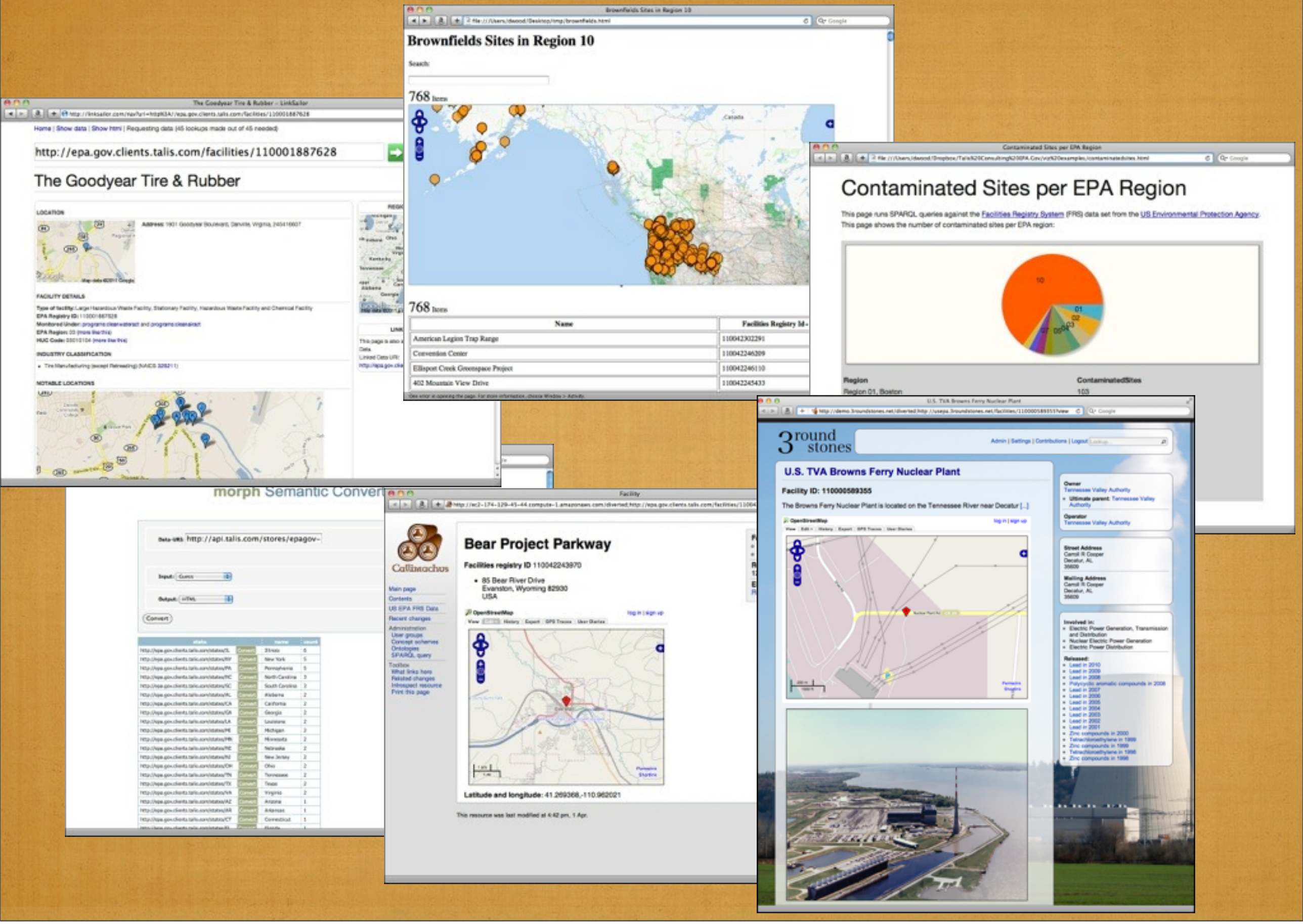
Data comes from DBpedia, WWF and other Linked Data sites. BBC curates data on the Web to ensure data quality. This is a fundamentally new business model for corporate data (and an indication of how far the EU/UK are in front of the US in this nascent market).

Monday, March 12, 12

Each HTML page is paired with a machine-readable data representation.

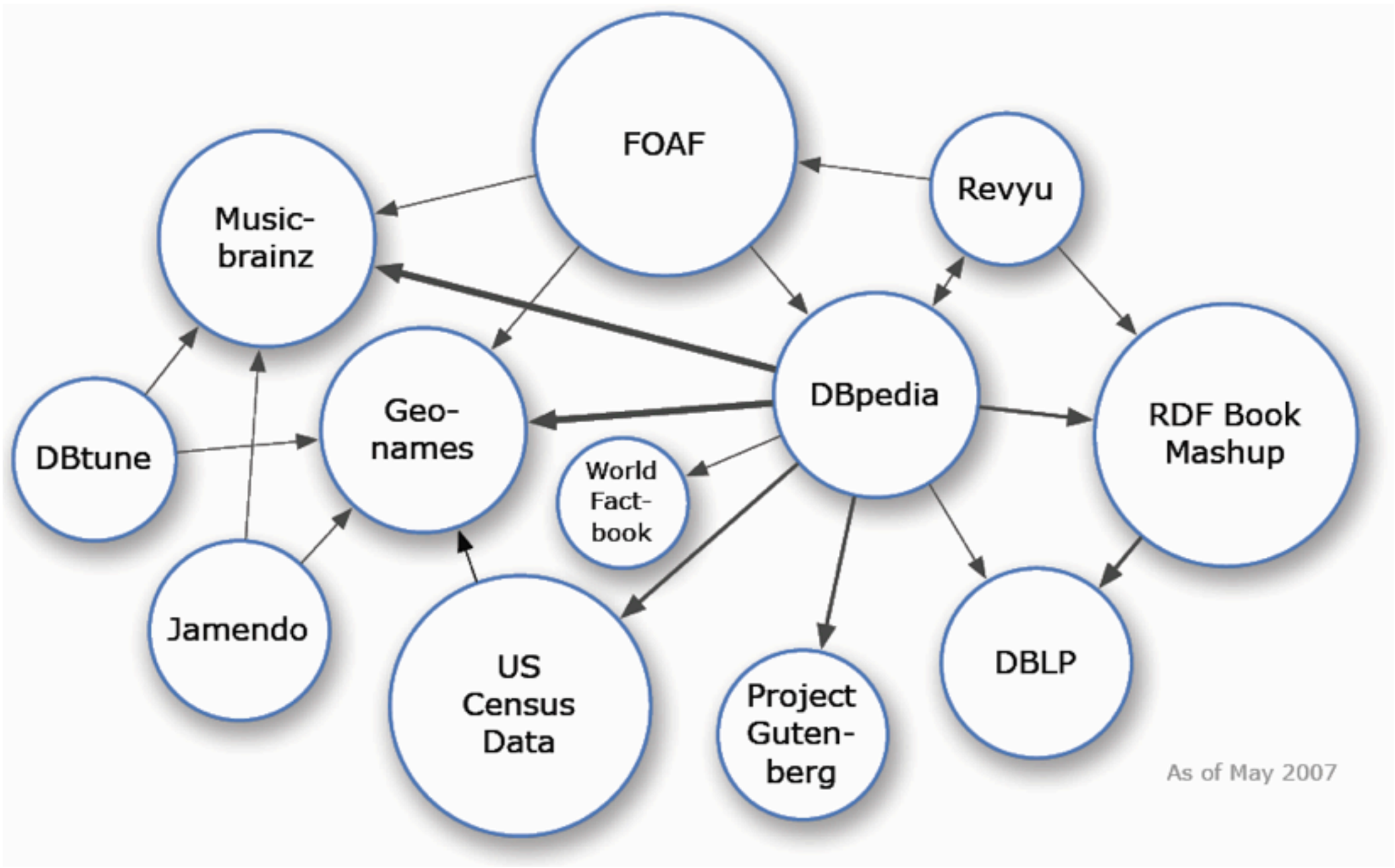
We've Seen This Before

Like HTML and RDF, credit cards have a human-readable side and a machine-readable side.

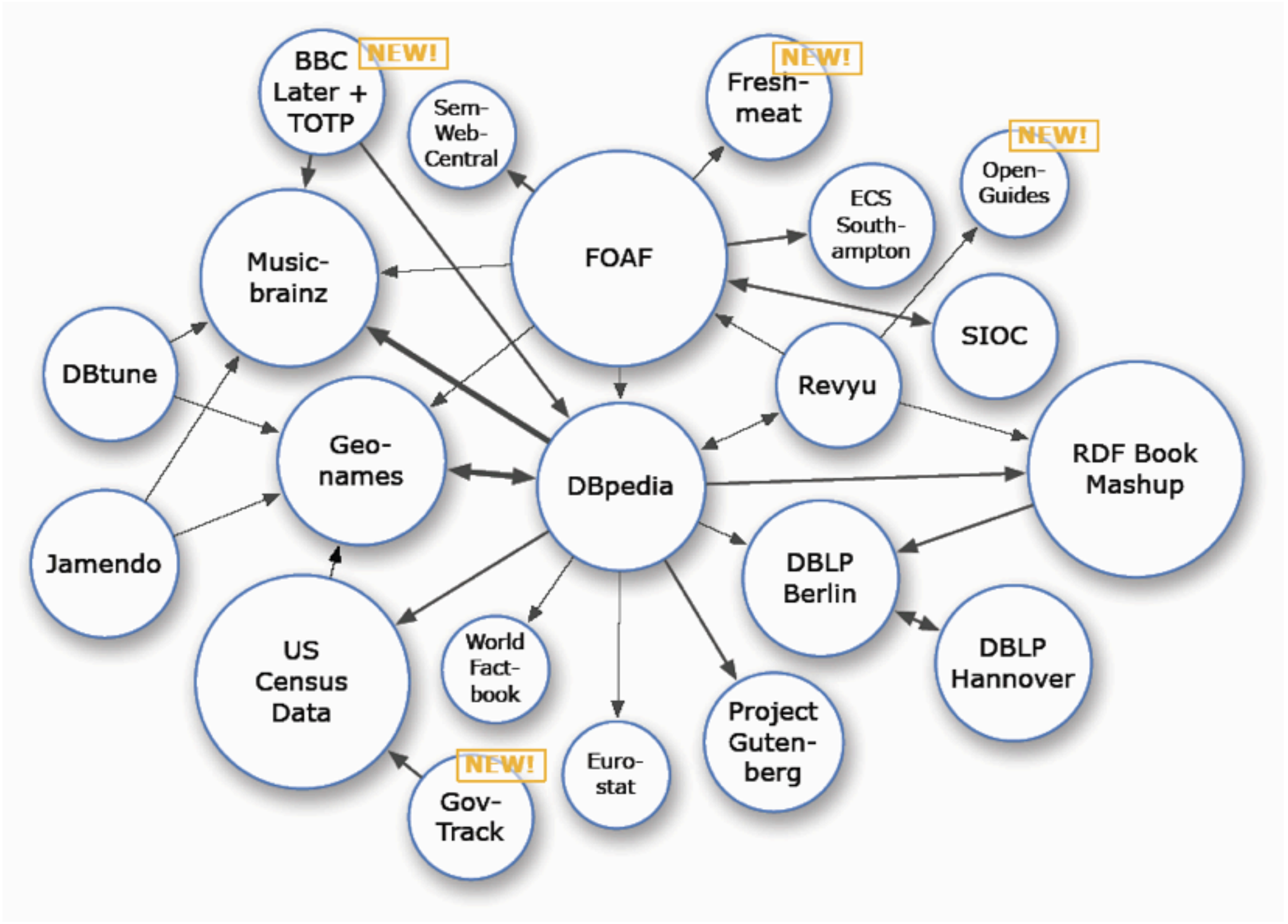


Monday, March 12, 12

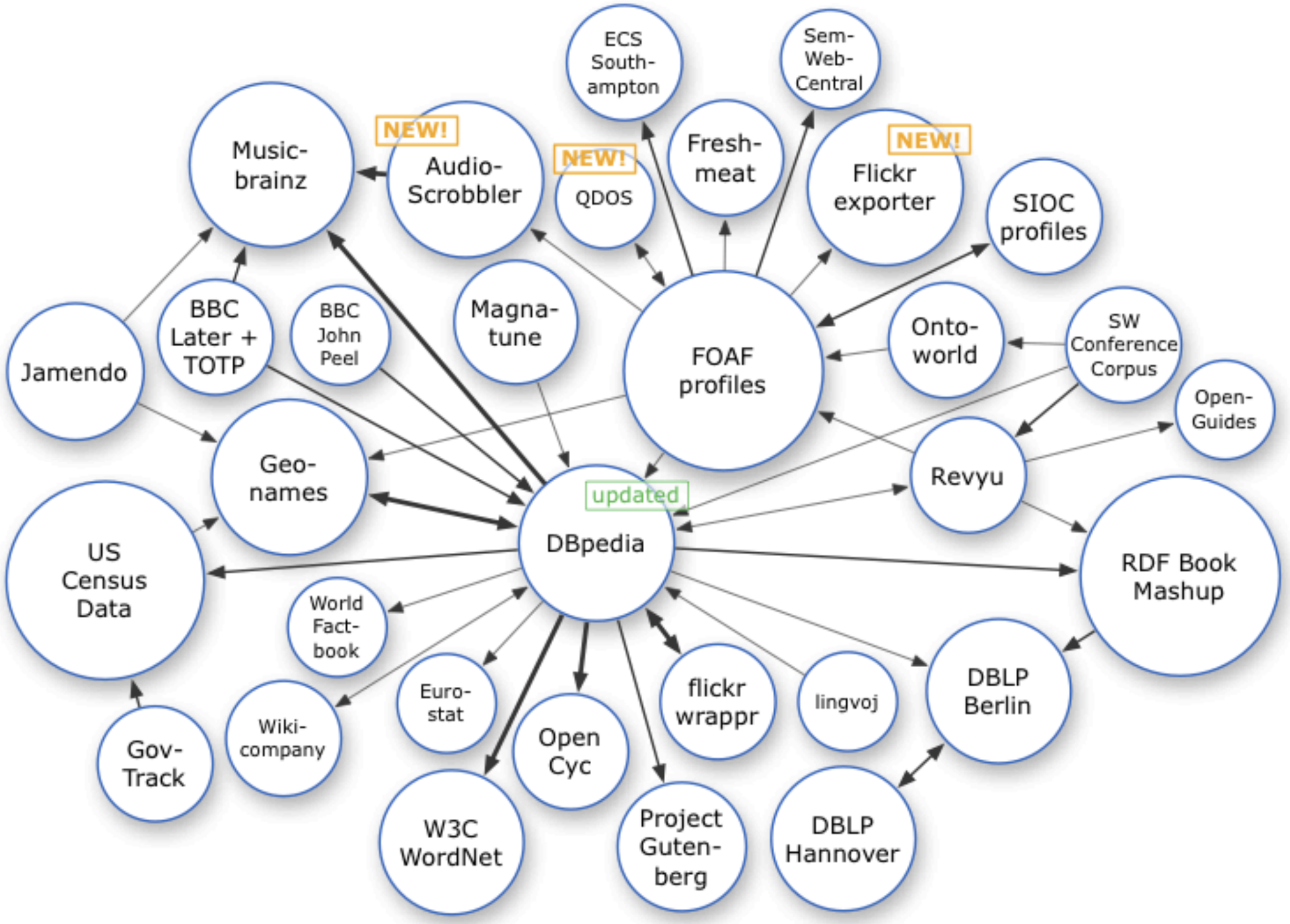
Linked Data may be rapidly repurposed, displayed in different ways for a variety of uses.
Challenge: What can you do by combining data?



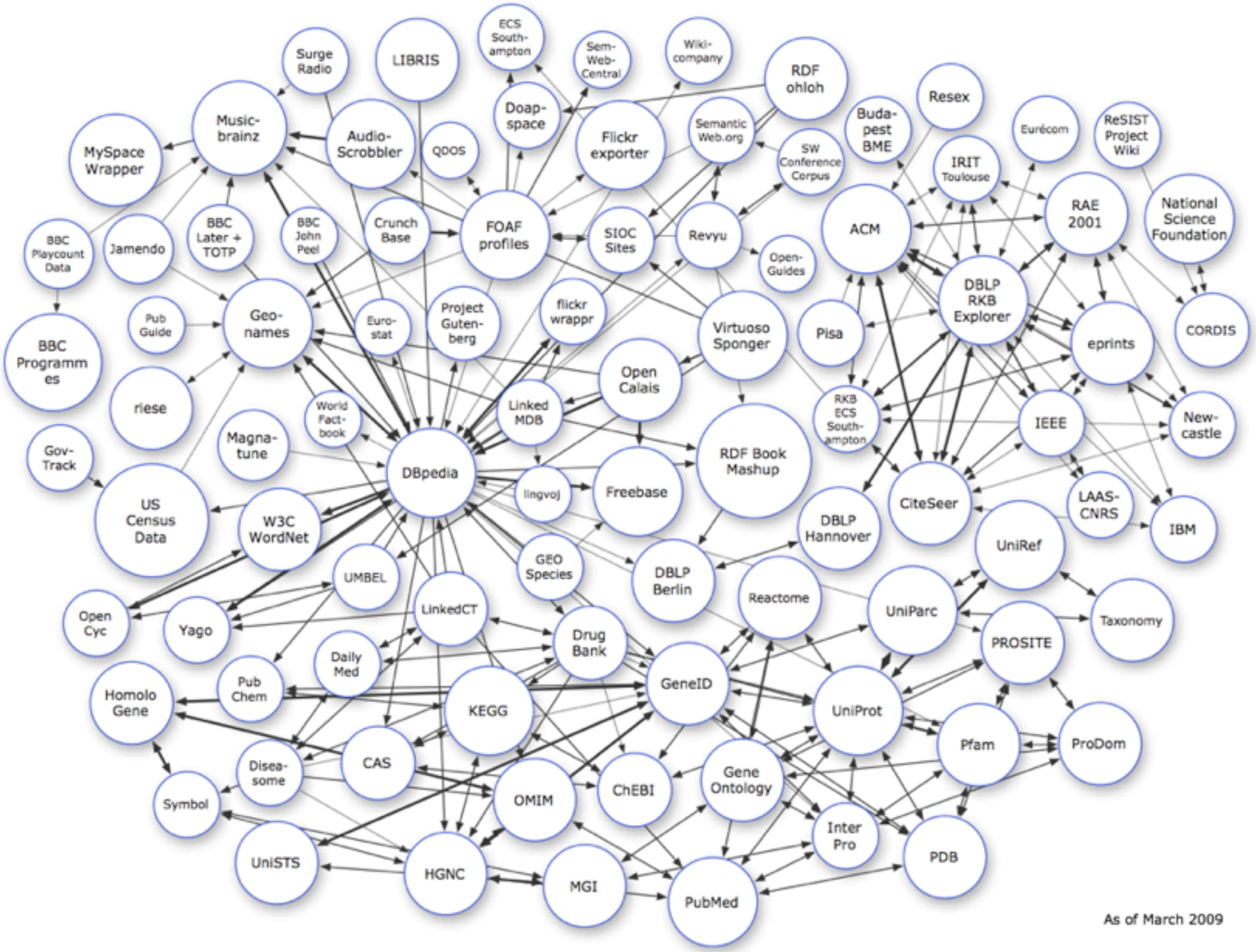
Some Linked Data is public. The Linked Open Data project consists of free and open data on the Web that has been interlinked together. Volunteers interlinked data that was already freely available.
May 2007



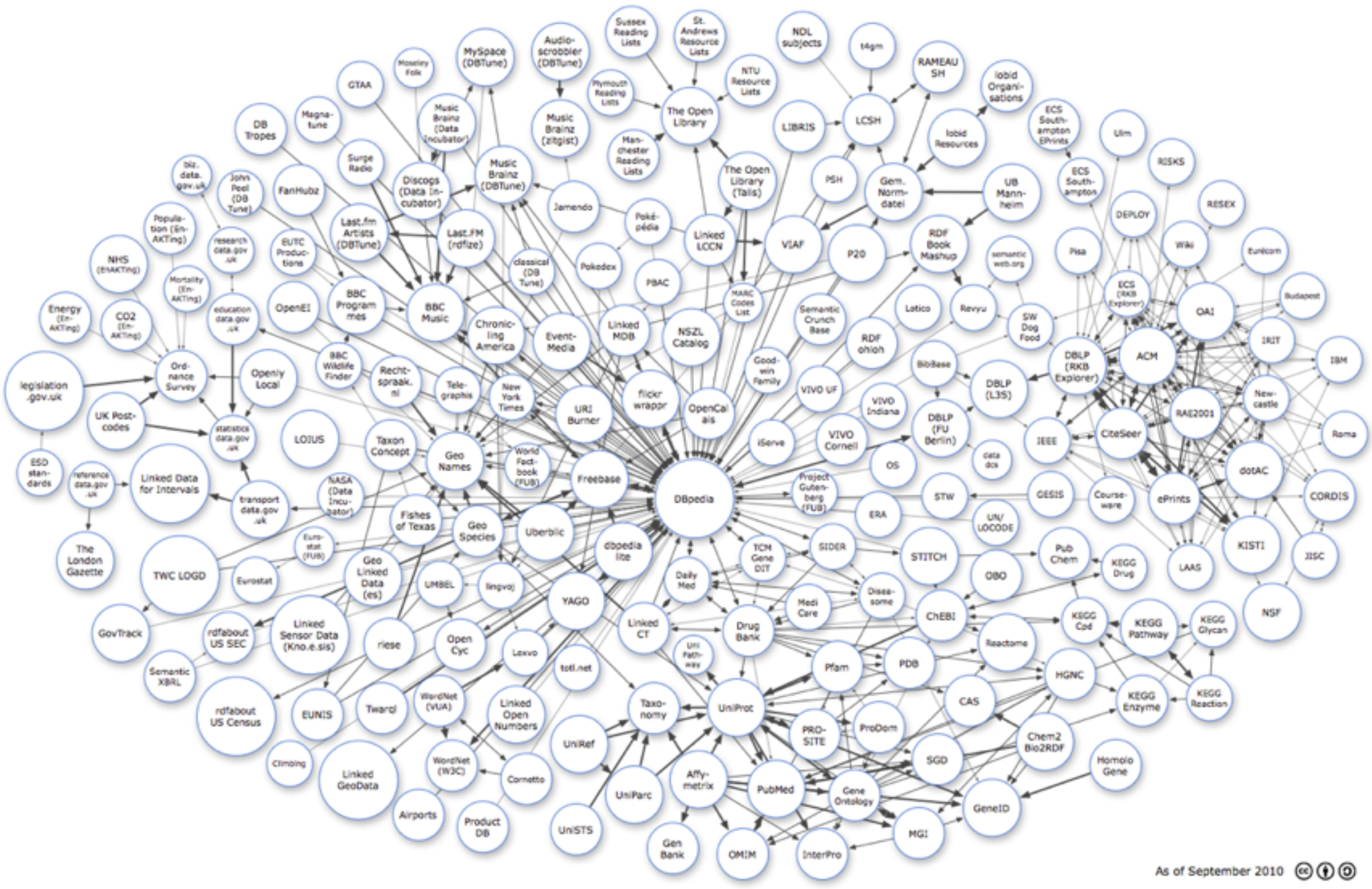
October 2007



February 2008

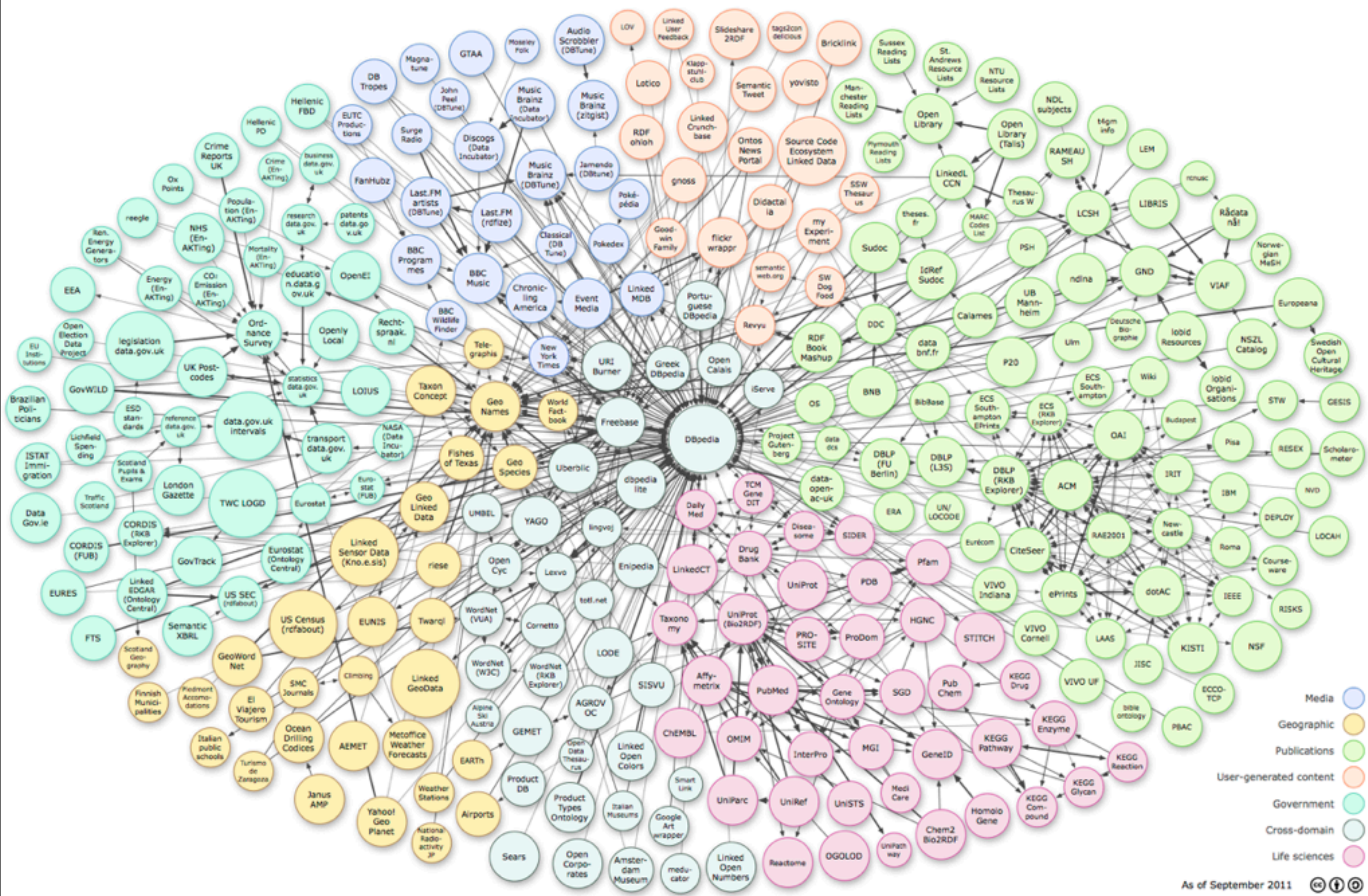

ECS South-ampton
Sem-Web-Central
Wiki-company
RDF ohloh
Surge Radio
LIBRIS
Doap-space
Flickr exporter
Semantic Web.org
SW Conference Corpus
Buda-pest BME
Resex
Eurécom
ReSIST Project Wiki
Music-brainz
Audio-Scrobbler
QDOS
IRIT Toulouse
MySpace Wrapper
BBC Playcount Data
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
SIOC Sites
Revyu
Open-Guides
ACM
RAE 2001
National Science Foundation
DBLP RKB Explorer
Pub Guide
Geo-names
Euro-stat
Project Guten-berg
flickr wrappr
Virtuoso Sponger
Pisa
eprints
CORDIS
BBC Programmes
riese
Open Calais
RKB ECS South-ampton
IEEE
New-castle
World Fact-book
Linked MDB
Gov-Track
Magna-tune
DBpedia
RDF Book Mashup
lingvoj
Freebase
CiteSeer
LAAS-CNRS
US Census Data
W3C WordNet
DBLP Hannover
UniRef
IBM
UMBEL
GEO Species
DBLP Berlin
Reactome
UniParc
Taxonomy
Open Cyc
Yago
LinkedCT
Drug Bank
PROSITE
Daily Med
Pub Chem
GeneID
Homolo Gene
KEGG
UniProt
Pfam
ProDom
Disea-some
CAS
Gene Ontology
ChEBI
OMIM
Symbol
Inter Pro
PDB
UniSTS
HGNC
MGI
PubMed
As of March 2009

March 2009



Monday, March 12, 12

September 2010

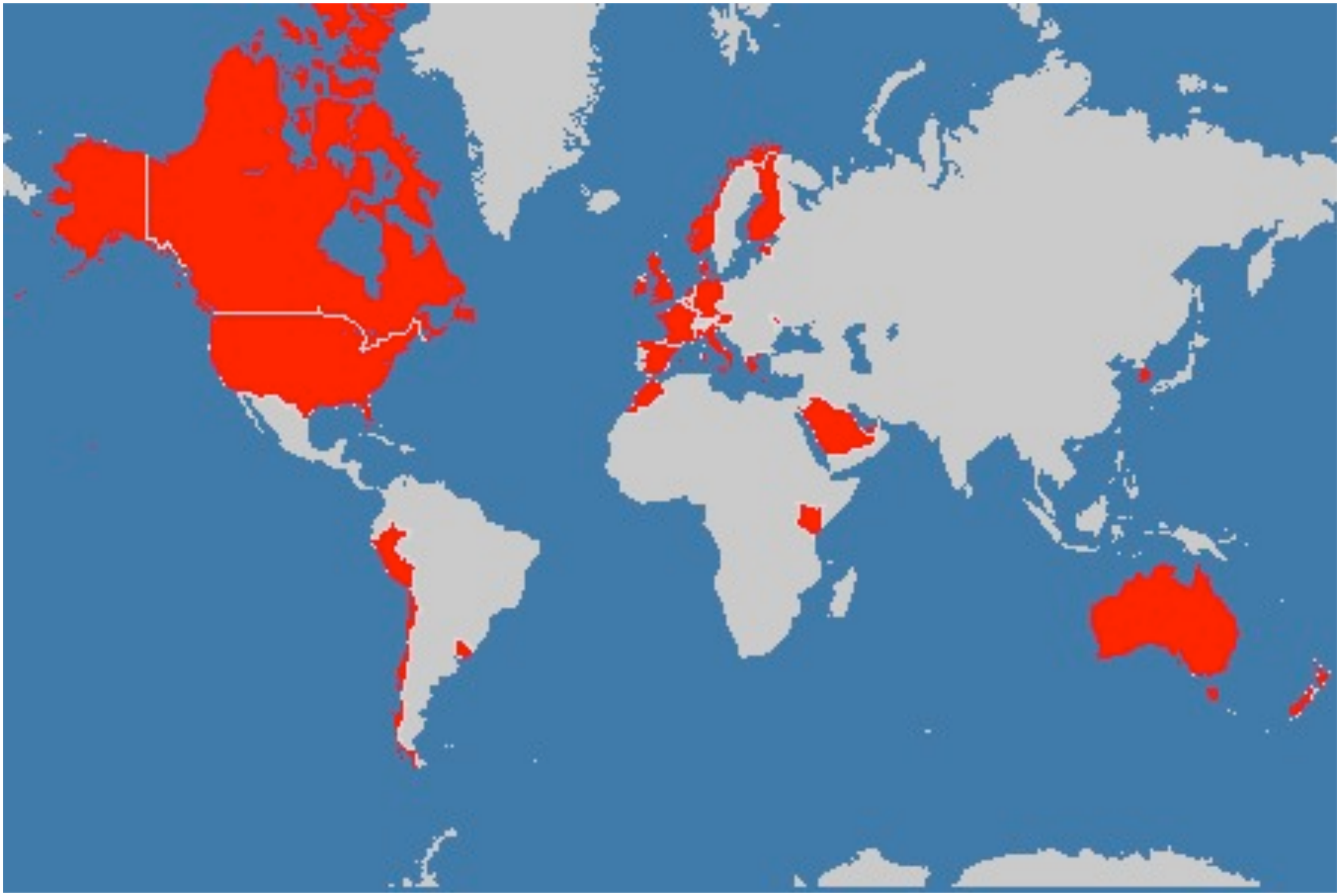


Monday, March 12, 12

September 2011: 295 datasets that meet the LOD Cloud criteria, consisting of over 31 billion RDF triples and are interlinked by around 504 million links.

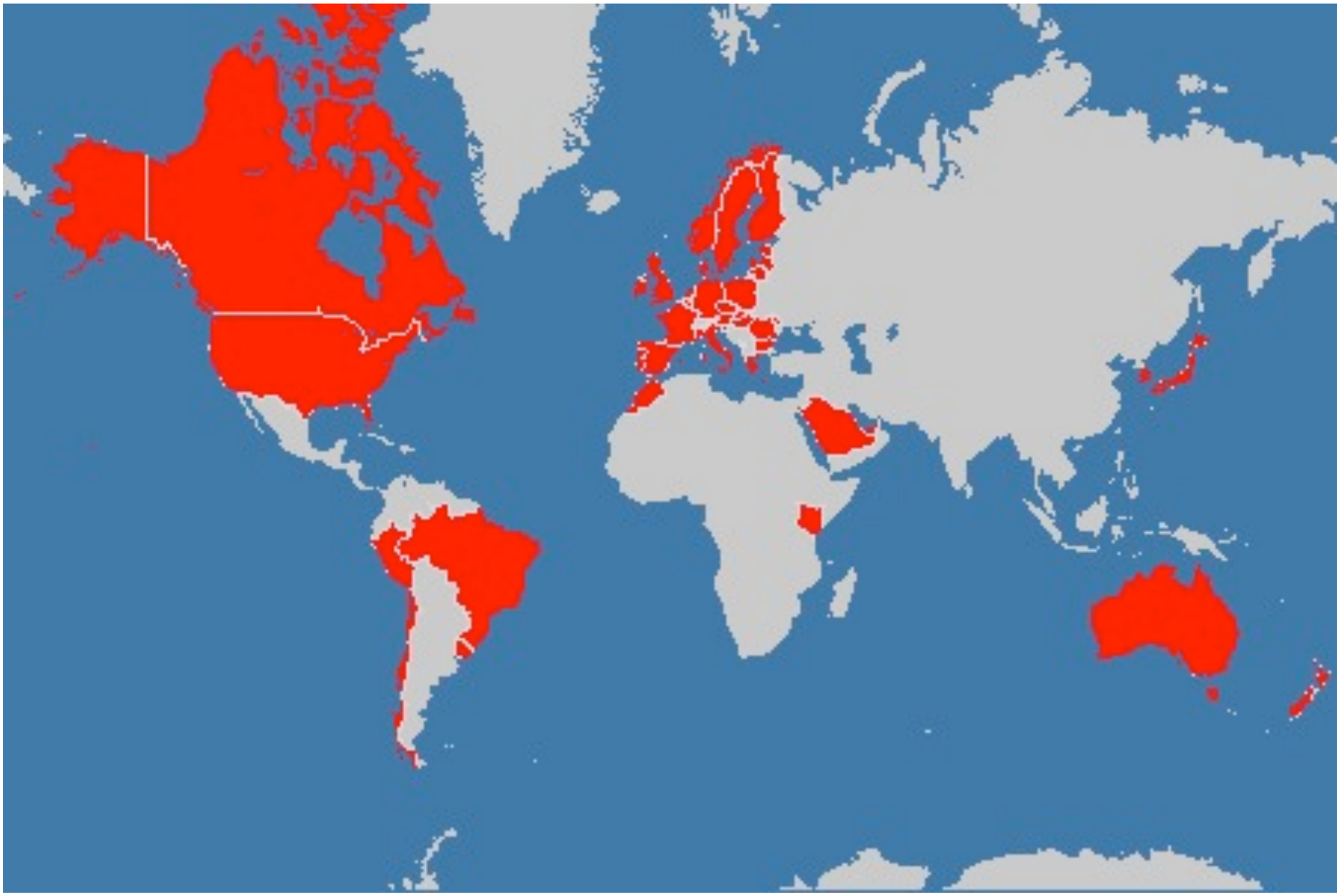
Why **Open** Government Data? Because everyone wants to be someone's hero. Politicians, civil servants, military officers, Facebook users, Twitter users... Opening our data allows others to use it - and give us credit.

Official Government Data in 2012



Open government data sites in 2011: 30 countriess (15%).
Source: http://www.data.gov/opendatasites
However, there are a variety of formats, little in the way of standards adoption.

...Adding in Third-Party Publishers



Add in LOGD published to CKAN.net, meeting LOD standards for license, size, interlinking, **regardless of publisher**: 28 countries (12%, some overlaps). The total becomes 48 countries (24%)

Note that EU publishes for Europe; UK publishes most.

Source: http://thedatahub.org/dataset

Largest Publisher of Government LOD

Most government LOD published to CKAN.net: UK
Source: http://thedatahub.org/dataset

Will Linked Data cross Geoffrey Moore's chasm? Some say that the Semantic Web will never happen, that relational databases are too entrenched, that no large organizations will ever invest in such an unproven technology.

YouTube	HDTV
watch videos	watch Better videos
Publish videos	
Share videos	
Rate videos	
Discuss videos	

Steven Johnson's excellent book "Where Good Ideas Come From" uses this example in the first chapter to encourage readers to think about why some technologies go viral and others don't. YouTube was an overnight success in two years, compared to the twenty it tool HDTV to take hold, in spite of significant governmental backing.

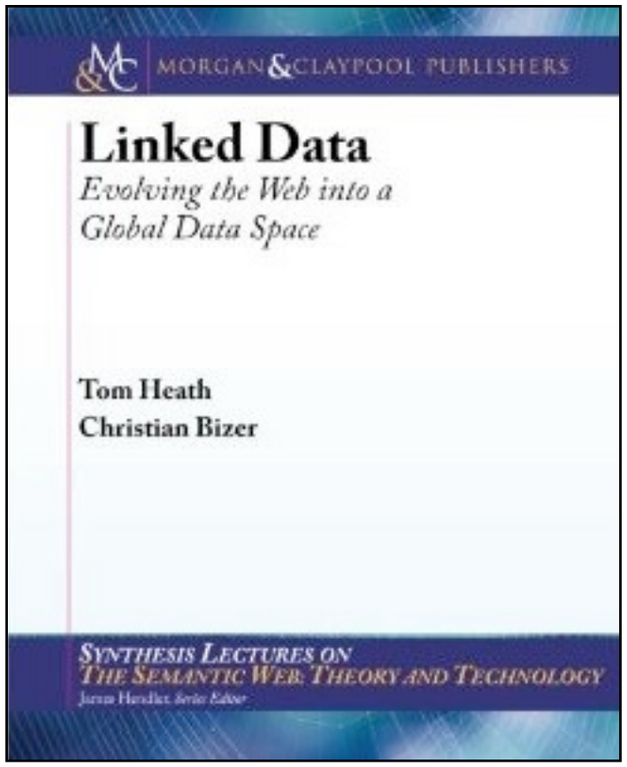
Linked Data	RDBMS
Use data	Use data
Publish data	
Share data	
Rate data	
Discuss data	

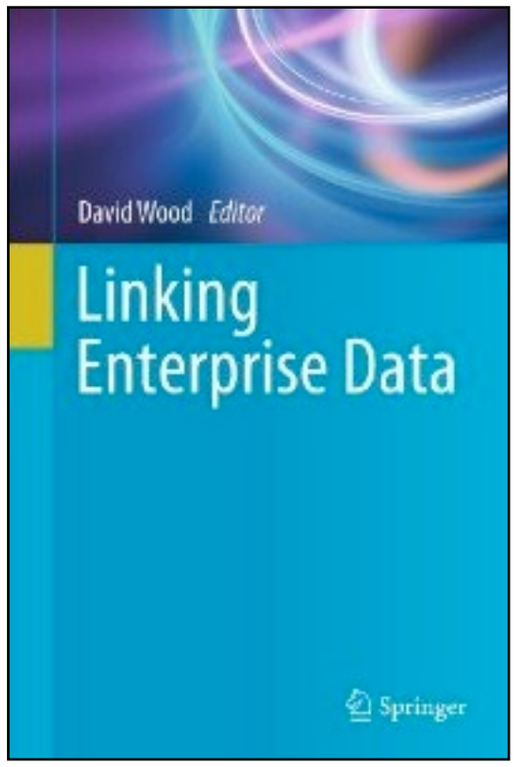
There is a reason that the Linked Open Data project has grown rapidly in four years, from a small volunteer effort to being used in production by governments and large companies.

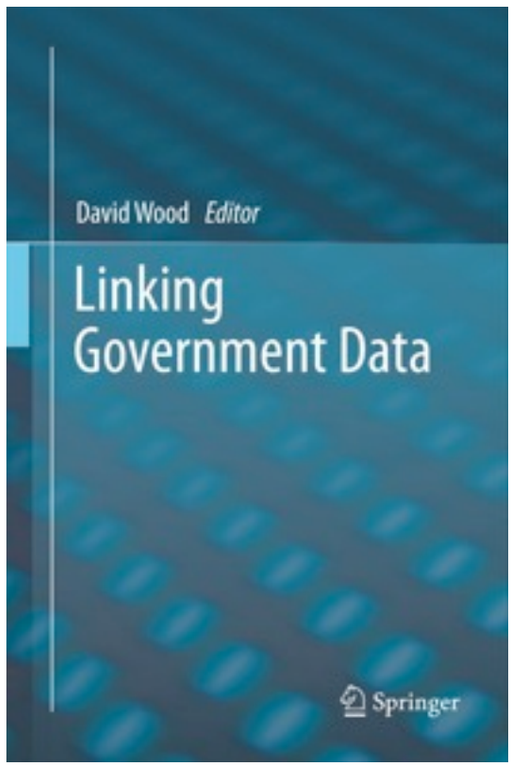
Monday, March 12, 12

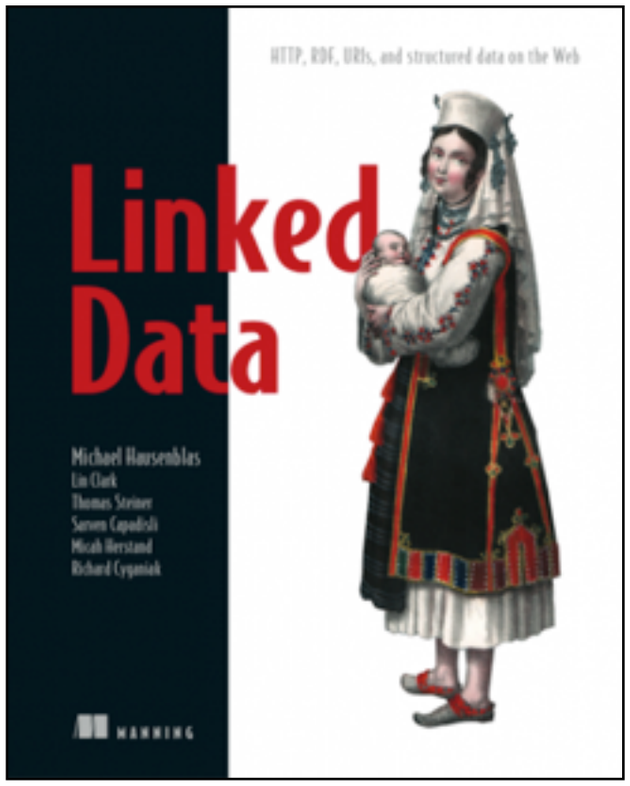
Some large commercial and government organizations would beg to differ...plus many more universities, small and medium sized companies and Open Source projects.
Many of these are careful, slow-moving entities who simply needed to find real solutions to real problems.

Further Reading









http://linkeddatabook.com/editions/1.0/

http://3roundstones.com/linking-enterprise-data/

http://3roundstones.com/linking-government-data/

http://www.linkeddatadeveloper.com/

http://linkeddatabook.com/editions/1.0/
http://3roundstones.com/linking-enterprise-data/
http://www.linkeddatadeveloper.com/

Credits

Country Maps (Use by license)	http://douweosinga.com/projects/visited
Mastering the Art of French Cooking cover (Fair use copyright)	https://en.wikipedia.org/wiki/File:MasteringTheArtOfFrenchCooking1edCover.jpg
Julia Child photo (CC-BY-SA licensed)	https://en.wikipedia.org/wiki/File:Julia_Child_restore.jpg
LOD Cloud Diagrams Richard Cyganiak, Anja Jentzsch, (CC-BY-SA)	http://lod-cloud.net/
Chasm Photo Travis S. (CC-BY-NC licensed)	http://www.flickr.com/photos/baggis/3860802929/
Corporate logos, Darkon Movie Poster, BBC screenshots, CAMC credit card image and book covers © their respective owners and used under Fair Use for educational purposes	
All other photos and drawings © 2010-11 David Wood, released under a CC-BY-SA license 	

3 round stones